CHAPTER 4: ANALYSIS OF FIELD TEST DATA

Introduction

As stated in Chapter 1, one focus of this preliminary report is the quality of the items that have been developed for use in the initial form of the HSEE. We have reviewed the process used in developing and editing new items and assembled two panels of educators who also reviewed the roughly 750 items that have been developed. (See Chapter 3 above.) In late May and early June, the development contractor (AIR) conducted a field test of these items, administering them to a reasonably large and representative sample of California 10th graders. As part of our evaluation, we have conducted independent analyses of the data collected from this field test. We describe these analyses and their results in this chapter.

Our analyses addressed three general issues. First, *what proportion of the items have good statistical properties?* The answer to this question provides an indication of the soundness of the development procedures and also will determine whether there are enough high-quality items to begin assembling one or more operational forms of the exam. Second, we want to provide a preliminary *assessment of the likely accuracy of scores* from the items that were field tested. Accuracy is one of the important issues that the State Board must consider as it reaches a decision about adopting the test in September or October. The third issue is whether we can estimate the *possible impact that the exam would have on different demographic groups* if it were administered today (or, more precisely, last Spring).

It is important to point out that our analyses are necessarily preliminary. We do not yet have results for the essay questions that are still being scored and we have not yet had a chance to review primary analyses being conducted by AIR. We plan to issue a supplemental report on August 25, after we receive additional information and conduct further analyses. Some of the counts and estimates in this preliminary report may change slightly in our final version or may differ slightly from results of analyses conducted by AIR. It seems likely that any such differences will be minor and will not affect the main conclusions that we draw from our preliminary analyses.

Field Test Design

Test Booklets

AIR constructed four test booklets (forms) of English language arts (ELA) items. Each form contained 100 multiple choice (MC) items followed by two constructed response (CR) essay items. A total of 59 different reading passages with MC questions (items) were tried out. Many of these passages were included in more than one test booklet with differences in some or all of the questions asked about the passage. The purpose of this repetition was to avoid asking too many different questions of any one student, but still allow the contractor to pick the best items for each passage when it is used in an operational form. In all, 338 unique multiple choice items were tried out in the field test, with 62 of these items included in two different forms (bringing the total MC items printed to 400, or 100 per booklet). Three different versions (subforms) of each booklet were created with the same MC items

but different CR items. In this way a total of 24 different CR items were tried out (2 for each of the 3 versions of each of the 4 forms).

AIR also constructed 4 forms of Mathematics (Math) items. Each form contained 99 MC items. There were no CR items for math and there was no overlap across the 4 different math booklets.

Field Test Sample

Details of the Field Test Sampling plan will be presented in AIR's report on the field test. Their basic goal was to ensure that the sample of students completing each test booklet covered a wide range of abilities and was generally representative of 10th grade students in California. The field test was not intended to provide normative information, as operational forms have not yet been assembled, so exact representativeness was not a primary concern. For each of the two exams, AIR sorted California schools by their level of performance on the corresponding 1999 STAR test and then picked 10 schools from the lowest performing tenth (decile) of these schools, 10 schools from the next lowest performing tenths, and so on up to 10 schools from the top performing 10th. This approach appears to be an effective way of obtaining samples of schools that span the full range of ELA and Math abilities.

For each of the selected schools, AIR requested up to 66 10th grade students. Some of the schools were too small to be able to comply with this request and others could not supply the total requested students because of end-of-year scheduling problems. Fortunately, AIR had been reasonably conservative in planning for this contingency and the resulting sample sizes appear adequate for most or all of the intended analyses. Within each school, the four different ELA or Math booklets were assigned to roughly one-fourth of the students tested. This provided "randomly equivalent" samples of students for the different booklets (the same ability levels except for random factors in the assignment to booklet that become negligible with large sample sizes).

Table 4.1 shows the total number of students completing each booklet. In these and the tables that follow, a small number of students with missing form codes or no valid item responses (86 in the ELA sample and 84 in the Math sample) were deleted from our analyses. Even though the tests were long, nearly all students responded to all of the items. Only 6% of the ELA sample and 5% of the Math sample failed to respond to (omitted) more than five of the 100 or 99 items.

Table 4.1 also shows the average "total correct" scores for the 100 (ELA) or 99 (Math) items presented to each student. For both subjects, these averages and the standard deviations (which show how much the scores varied across different students) were very similar across the four test forms. Assuming the random assignment of students to booklets worked as intended, this similarity in number correct scores suggests that the items in each of the different booklets were of comparable average difficulty.

Table 4.1 Average Total Correct Scores by Subject and Field Test Form

Subject	Form	Sample Size	Average (%)	Standard Deviation
ELA–MC	1.x	998	58.9	20.6
	2.x	1017	59.3	20.8
	3.x	906	58.9	21.0
	4.x	836	58.5	19.6
	ALL	3757	58.9	20.5
Math	1	1009	46.1	19.2
	2	922	47.3	18.0
	3	1020	48.1	17.9
	4	969	46.7	17.0
	ALL	3920	47.1	18.1

Item Difficulties

The results in Table 4.1 above also provides important information on the average difficulty of the HSEE items for California 10th Graders. Each ELA form had 100 MC items, so an average score of 59 means that, across items, the average percent answering correctly was also 59. These items were all 4-option multiple-choice items. Because of guessing, the percentage of students answering correctly is greater than the percentage that actually know the right answer. For example, suppose 45.3% of the students knew the correct answer and the other 54.7% guessed randomly. All of the students knowing the answer (45.3%) would answer correctly and one fourth of the students who did not know the answer (13.7%) would answer correctly through random guessing, so the expected percent answering correctly would be 59% (45.3 + 13.7). This example suggests that these are relatively difficult items, with fewer than half the students knowing the answer for the average item.

For Math the items appear even more difficult. An average score of 47.05 on a 99-item test translates into an average of 47.5% correct on the average item. This result would be obtained if only 30% of the students knew the correct answer and one-quarter of the other 70% (17.5%) answered correctly through guessing. Thus it seems that less than one-third of the students knew the correct answer for the average math item.

We also examined the distribution of number correct scores for different demographic groups as shown in Tables 4.2 and 4.3. These results provide a preliminary indication of the relative difficulty of the HSEE items for different groups of students. In a later section of this chapter, we will use this information, along with data from the 1999 STAR administration, to provide very preliminary suggestions about the possible differential impact (passing rates) of the HSEE for these different groups of students.

Table 4.2 Average Total Scores by Gender

Subject	Gender	N	Mean (%)	Standard Deviation
ELA–MC	Female	1835	62.9	18.9
	Male	1895	55.2	21.2
Math	Female	1915	46.8	17.3
	Male	1988	47.4	18.8

Table 4.3 Average Total Scores by Race and Language Fluency

Subject	Race/Language Status	N	Mean (%)	Standard Deviation
ELA–MC	African American (1)	197	50.2	20.0
	Asian (3)	266	68.8	18.6
	Hispanic (5)	1314	50.9	18.3
	White (7)	1610	65.7	19.4
	L.E.P	430	40.5	14.4
Math	African American (1)	300	41.0	15.4
	Asian (3)	318	57.6	19.8
	Hispanic (5)	1108	38.6	13.8
	White (7)	1800	52.0	18.1
	L.E.P	318	35.6	14.8

Item Screening

A total of 396 Math items were organized into 4 field test booklets of 99 items each. Four test booklets of ELA items were also included in the field test. Each of the ELA booklets contained 100 multiple choice (MC) items. Each booklet was further divided into 3 "subforms" with two distinct constructed response items in each subform. Across the forms and subforms, a total of 24 constructed response items was administered to about 350 students each. Scores for these items are not yet available, so their analyses will be included in a supplemental report.

The 100 MC items included in the 4 forms were not all unique. There were 338 unique MC items with 62 of these items included in two different forms each to provide a basis for linking statistics for item in the different test forms. In our item screening analyses, we used statistics from the first occurrence (lowest form number) of each of these duplicated items. To the extent feasible, our supplemental report will include analyses of differences in statistical results across forms for these duplicated items and alternatives for pooling results across forms.

We made a very preliminary effort to estimate the number of field test items with statistical properties that suggest they would need to be dropped or revised (and re-tested) before being used in operational forms. Statistical indicators were used to assess: (1) whether items were inappropriately easy or difficult, (2) whether the item score provided information that was at odds with (did not generalize to) the information provided by the other items, and (3) whether the item appeared to function differently for different demographic groups (females, Hispanics, or African Americans).

Item Difficulty

We computed the percent passing (p-values) for each item. In subsequent analyses, it might be possible and desirable to adjust these p-values for differences between the field test samples and the total population of California's 10^{th} grade students. As noted above, the procedures used in drawing the sample should have been sufficient to ensure that any such adjustments would be minor. Item difficulty screens are used to weed out items, which, although they could be perfectly valid, provide little or no useful information. More often than not, extreme item difficulties also reflect item flaws so that most of the items screened out are not valid measures of the intended standards as well as being inefficient. For

example, if nearly all students pass an item, it may well be that the distracters (incorrect options) are not plausible or that something in the item text "gives away" the correct answer. Similarly, if the percentage answering correctly is at the guessing level (suggesting that no one really knows the correct answer), the item provides little information about student skills and is likely to be flawed. In this case, the item could be incorrectly keyed or have no correct option or have some problem in the text that leads even able students astray. We flagged items with passing rates above 95% as too easy and those with passing rates below 25% (the guessing level for 4-option items) as too difficult.

Item-Total Correlation

Another indicator of potential item problems is when results from the item disagree with (fail to generalize to) the scores on other items. The item-total correlation coefficient measures the extent to which students who answer the item correctly also score well on the rest of the test. Because the item score is dichotomous (scored pass or fail) and the total score has a continuous (more normal) distribution, the range of the item-total correlations is limited, particularly when the percentage of students passing the item is much different from 50. We computed a Clemans-Brogden biserial correlation coefficient (Lord & Novick, 1968, page 341) that corrects for differences in item difficulty. Possible values range from –1.0 to +1.0 with positive values indicating agreement between the item score and the total score. We flagged all items with values less than 0.2 as having a generalizability problem. Often these items are mis-keyed or have ambiguities in the text or options that limit their validity as a measure of achievement of the targeted standards.

Differential Item Functioning (DIF)

It is common practice to look for differences in the way an item functions across different groups of students. In most analyses of differential item functioning (DIF), a focal group is identified that is of specific concern. The rates at which members of this group answer an item correctly (pass) are compared to passing rates for a second reference group. In our analyses, Hispanics, African Americans, and Females were the focal groups of interest. In each case, statistics for these students were compared to statistics for all other students in the field test.

The issue is not just whether there are different passing rates for these different groups. The question addressed in DIF analyses is whether group differences in passing rates for some items are significantly larger than the differences in passing rates for the other items. Another way of framing this issue is to ask whether students from different groups who are at the same overall level of achievement (usually indicated by the total test score) have the same probability of answering the item correctly.

We computed DIF statistics[6] for females, Hispanics and African Americans–the groups of most common concern in test bias studies. The sample sizes for females and Hispanics (more then 400 and 300 per test form respectively) were large enough to detect moderate and large DIF reliably. The sample size for African Americans was much smaller, generally 40 to 50 per item. Only a few items were flagged as having potentially significant DIF for this group, in part because the sample size was not large enough to allow detection of items with only moderate DIF.

Note that a finding of significant DIF does not necessarily mean that an item is not a valid measure of the intended standard. Group differences in preparation can lead to greater group differences on some items than on others. For example, suppose that male and female students took algebra at the same rate, but many more male students went on to take geometry by the 10th grade. We would expect larger gender differences in passing rates for geometry items than for algebra items, even if all items were perfectly valid measures of their intended content. A common practice is to flag all items with significant DIF values for further content and sensitivity review. Many of these items would then be subsequently accepted and used without further changes. We used a relatively high cut-off (the .01 level) to estimate the proportion of items that would eventually be screened out because of DIF concerns.

Item Screening Results

Table 4.4 summarizes our item screening results. It should be noted that these are preliminary estimates based on statistical criteria only. AIR will end up with somewhat different results using somewhat different statistical criteria and incorporating editorial, as well as statistical, review of flagged items.

Overall the results show the tests to be highly effective in correlating items to standards and asking questions correctly. In many programs, half of the items or more are screened out on the basis of initial field test results. We flagged only 1 out of 4 of the Math items and 1 out of 8 of the ELA items. The very high survival rates for the HSEE items shows a high degree of effectiveness in the item development and review procedures and reflects the fact that some of these items have been previously screened.

[6] A commonly used DIF statistic, the Mantel-Haenszel log odds ratio (Mantel & Haenszel, 1959), compares the odds of passing the item (percent correct/percent incorrect) for focal and reference group members at each different total score level. An odds ratio is computed for each total score level (indicating comparable overall ability). If the odds of passing for the focal group are the same as for the reference group, the ratio of the odds values is 1.0 and the logarithm of this ratio is 0.0. To the extent that the log-odds values (across all of the score levels) are different from 0.0, the item is said to function differently (be disproportionately hard or easy) for the focal and reference groups. We computed a chi-square statistic (see Dorans & Holland, 1993, page 40) that tests whether the Mantel-Haenszel statistic is different from 0.0. We flagged cases where the statistic was greater than 7.8794. This corresponds to the .005 level for a one-degree chi-square, meaning that there was less than .01 chance of getting a value this large (or a correspondingly small one) by chance alone.

Table 4.4 Percent of Items Screened Out by Various Statistical Criteria

Subject/Statistic	ELA-MC	Math
Total field test Items	338	396
Number passing all screens	294	307
Percent passing all screens	87.0%	77.5%
% Too easy*	0.0%	0.0%
% Too hard	2.1%	7.6%
% Low Item-Total Correlation	3.6%	6.6%
% DIF–Female	7.4%	9.6%
% DIF–Hispanic	3.0%	0.8%
% DIF–African American	0.0%	1.0%

* Note: Percents add to more than 100 because some items were flagged for more than one reason.

Potential Test Accuracy

The second issue that we sought to address in our preliminary analyses of the field test data was how accurate HSEE forms might be. This will necessarily be a technically dense discussion. Accuracy involves concepts of each student's "true" score (generally defined as the average of the scores across an infinite number of parallel forms) and of error (the difference between the score from a single testing and this true score). Models and estimates of the relative size of errors in test scores are used to construct confidence bounds in score reports, similar to the "margin of error" figures now commonly reported for political polls.

Most of our analyses here are based on models that come from Item Response Theory (IRT). We provide a brief description of the key IRT concepts that underlie these analyses in Appendix A. Readers are referred to more standard texts (Lord & Novick, 1968; Lord 1980; or Hambleton and Swaminathan, 1985) for a more detailed presentation.

Method

Simulated Test Forms. Table 4.5 shows Coefficient Alpha reliability estimates for each of the field test forms. These values are all quite high, which is not surprising given the length of the tryout booklets (which mirrored the target length for operational forms). Each of these tryout booklets contained some items that are likely to be screened out, so further analyses of the accuracy of these particular forms was not judged to be useful.

Table 4.5. Coefficient Alpha Reliability Estimates for each Field Test Form

Form	ELA	Math
1	.96	.95
2	.96	.94
3	.96	.96
4	.95	.94

We used the MULTILOG program (Thissen, 1991) to estimate IRT item parameters for each of the field test booklets. MULTILOG was chosen because it will handle CR items with more than 2 score levels together with MC items scored dichotomously. For items with negative item-total correlations, we fixed the parameters in advance (setting the slope to 0 to estimate a flat-line ICC). Each of the runs "converged" on the first try and there were no extreme values for the parameter estimates (other than for the "fixed" items).

In our further analyses of test score accuracy, we created "pseudo-forms" by selecting two (for ELA) or three (for Math) distinct sets of items from the pool of 294 ELA items and 307 math items with no statistical flags.[7] Note that, except for screening out questionable items, item selection was based on content classification only and not on any statistical properties. Specifically, there was no attempt to match item difficulties across forms as there would be in selecting items for operational test forms.

Simulated Examinees. For each of the pseudo-forms, we computed test accuracy information for 100 different levels of ability (simulated examinees). The IRT item parameter estimates were based on the assumption that there was a normal distribution of achievement. We sliced the area under the normal curve into 100 equal-area bars (each bar thus representing 1% of the examinee population). We took the midpoint on the ability scale for each slice as the level of ability we wanted to analyze. These levels of ability thus correspond generally to percentiles, except that percentiles are defined by the boundaries between each slice rather than the midpoints.

Simulated Decision Points. While recommendations for test content appear close to final, there has been little specific discussion, let alone any recommendations, of how well students will be expected to perform to pass each of the two tests. Normative information from STAR and NAEP suggest that California students are further behind in reading than in math. The field test data, however, show that the items assessing the Math standards appear to be more difficult than the items assessing the ELA standards.

The decision about how high students must score to pass these tests is a policy judgment. There are no right or wrong judgments (although there may be incorrect interpretations of the results), and it is not possible to predict, in advance, what the final decisions will be. There are, however, common conceptions based on the idea that a pool of items represent evenly some target domain that would suggest some plausible decision points. In less formal testing, it is common to think that about 70% correct is passing. No one can be expected to master 100% of any domain of knowledge or skill, but something above half of the domain would be reasonable. Seventy percent is between those two extremes.

Expecting students to answer 70% of the current HSEE items correctly would appear to be a reasonably high standard, given the difficulties estimated for the field test items. We used 70% correct as the upper end of the range of plausible passing standards and chose 50% to anchor the lower end. If the passing level were set much below 50% of the current items, it might be important to go back to the drawing board and develop some easier items. We used 60%-correct as an intermediate decision point in some of our analyses.

Results

Potential Percent Passing. For each of our 100 simulated examinees (ability levels) we computed their expected number correct score on each of the simulated forms. Figures 4.1

[7] To identify items for each pseudo-form, we grouped all of the items by the specific standard they were designed to test and then selected every "kth" item, where k was 294/100 for ELA and 307/99 for Math. This approach ensured as complete coverage of the different standards as possible. We used different starting points for each of the two ELA and three math pseudo-forms so that there would be no overlap in these forms.

and 4.2 show how the expected number correct scores would vary across the different forms and ability levels.

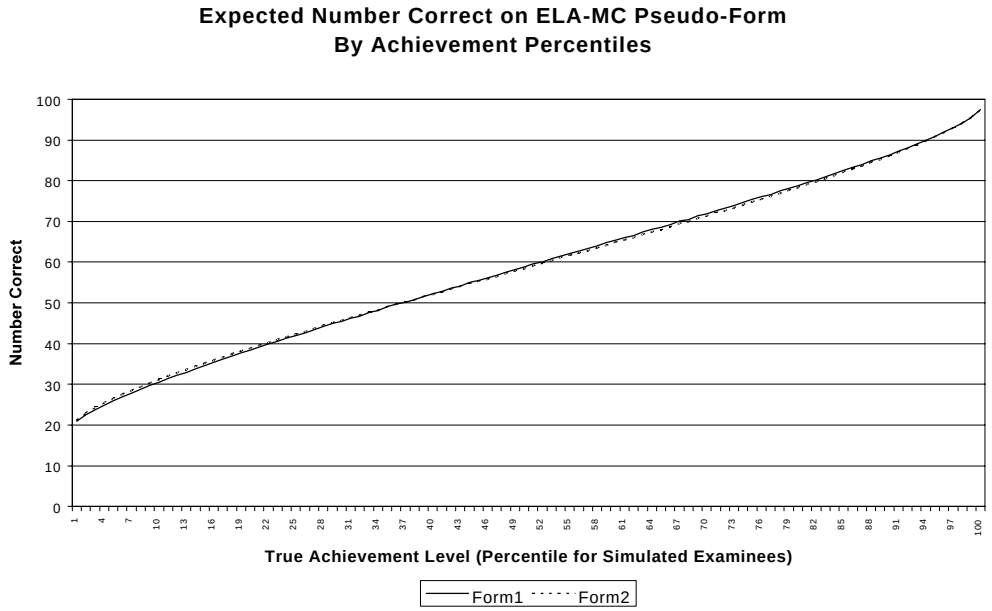

Figure 4.1 Expected number correct for each simulated examinee on each ELA pseudo-form.

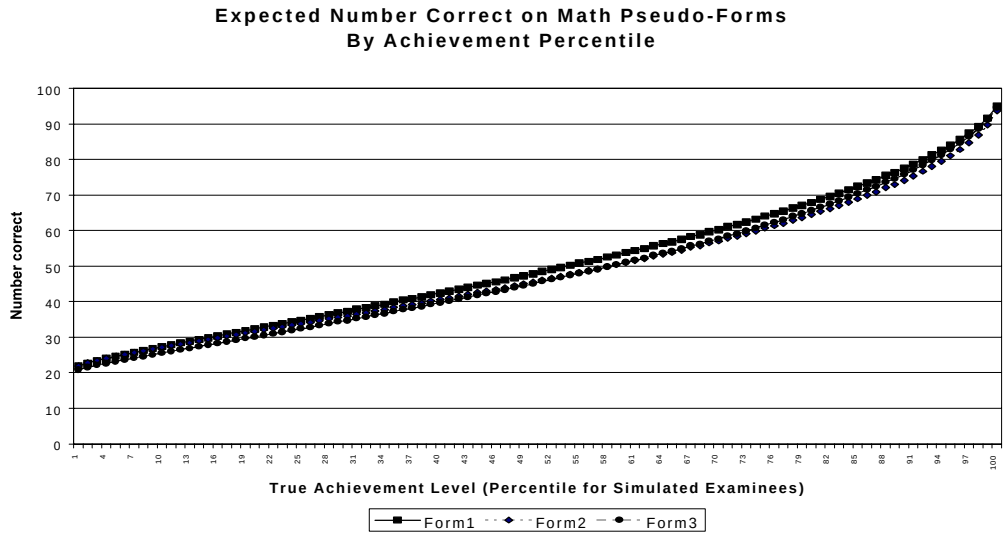

Figure 4.2. Expected number correct for each simulated examinee on each math pseudo-form.

Table 4.6 shows the percentage of students who would be expected to fall in the score ranges defined by decision points at 50, 60, and 70 out of 99 or 100 items correct. For ELA,

about one-third of the students would score below 50, one-third between 50 and 70, and one-third above 70. For Math, more than half of the students would score below 50, while only about 16% would pass if a total score of 70 were required. There was close agreement across the different pseudo-forms, even though no attempt was made to balance their statistical properties. Consequently, we pooled results across pseudo-forms in the remaining analyses.

Table 4.6 Percent of Simulated Examinees Scoring at Different Levels For Each ELA and Math Pseudo-Form

Subject / Pseudo Form	Percent of Students Expected to Score:			
	0–49	50–59	60–69	70 & above
ELA/1	36%	15%	15%	34%
ELA/2	36%	16%	16%	32%
ELA–average	36%	15.5%	15.5%	31%
Math/1	53%	16%	13%	18%
Math/2	58%	16%	12%	14%
Math/3	58%	15%	11%	16%
Math–Average	56%	16%	12%	16%

Decision Error. In characterizing the importance of test accuracy, we looked at decision errors as the most critical concern. Decision errors are made if a student passes the exam even though his or her true achievement level (defined on the expected number correct scale) is below the decision point (false positives) or if a student fails when his or her true achievement level is above the decision point (false negatives). Given the relatively low proportion of students expected to reach scores of 70 or higher, we chose 50 items correct as the decision point to focus on.

For each true achievement level, we computed the probability (based on the Item Response Theory (IRT) parameter estimates) that a student at that level would score below the decision point. Figures 4.3a and 4.3b show the proportion of time students would be expected to score below (fail) for each true achievement level.

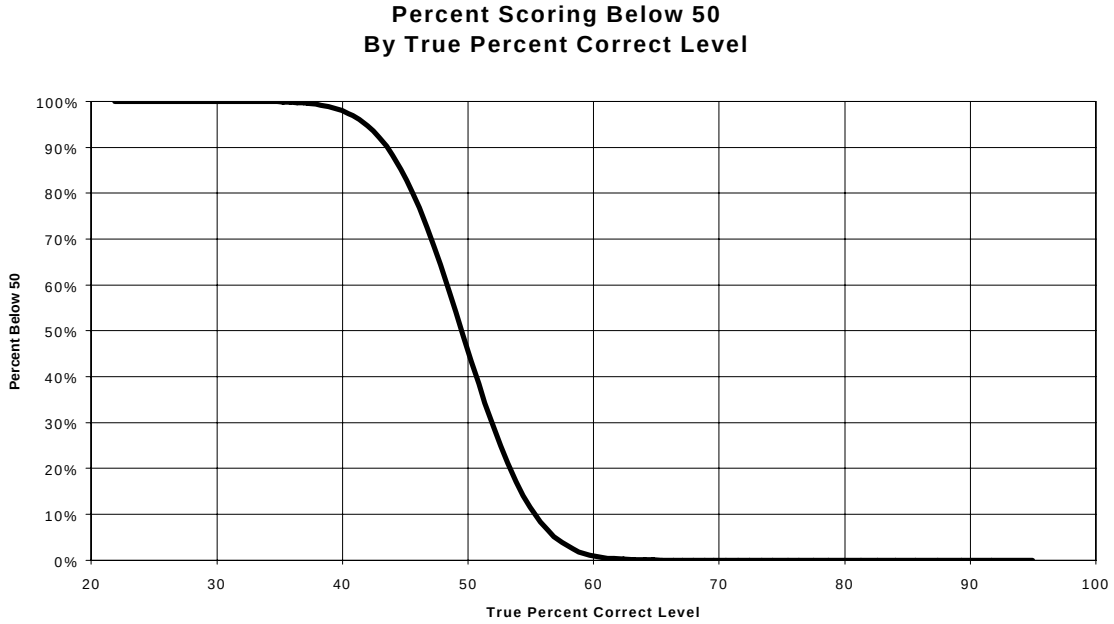

Figure 4.3a Percent Scoring Below 50 for Each True Achievement Level: ELA

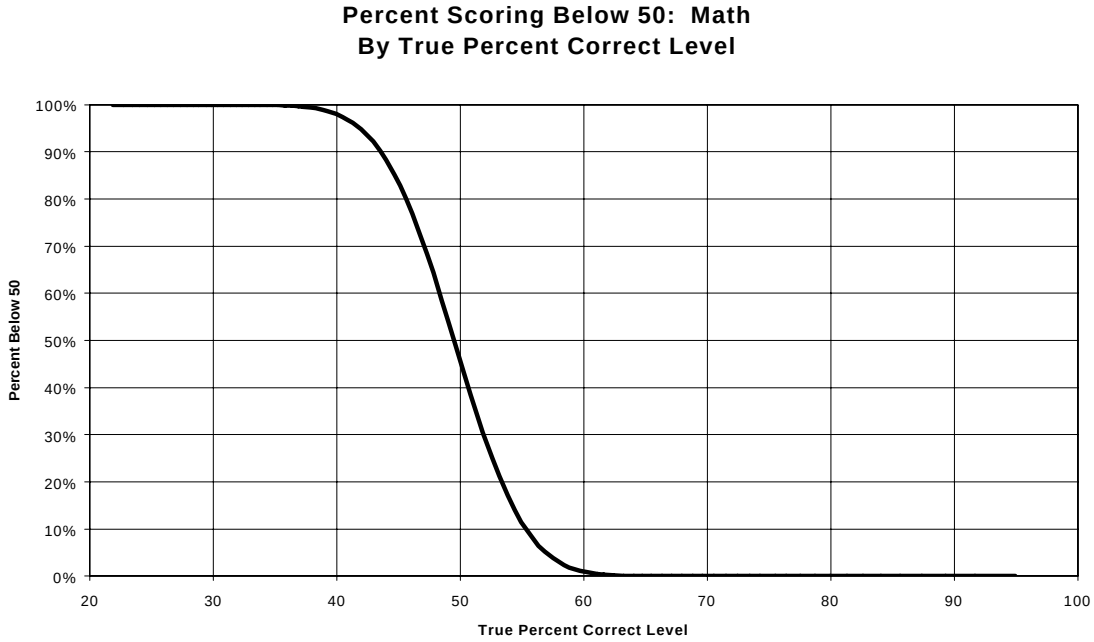

Figure 4.3b Percent Scoring Below 50 for Each True Achievement Level: Math

Nearly all students whose expected score is below 40 will fail and all students whose expected score is above 60 will pass. Between 40 and 60 there is a significant possibility of decision error. Of course, a student whose expected score is right at the decision point will pass or fail about half the time.

Table 4.7 summarizes the percent of the students in each expected number correct range (true achievement) whose single-test scores would fall in each of the decision-point ranges. Fewer than 9% of the students with expected scores below 50 would score in the ranges above 50 and fewer than 8% of the students with expected scores above 70 would score below 70. There is considerably greater uncertainty for students with expected scores between 50 and 70.

Table 4.7 Distribution of Observed Scores From a Single Testing for Students With Different True Scores (Expected Average Across Parallel Forms)

Subject	True (Expected) Number Correct	Percent of Students Who Would Actually Score:			
		0–49	50–59	60–69	70 & above
ELA	00–49	91.4%	8.5%	0.1%	0.0%
	50–59	16.0%	63.9%	19.8%	0.2%
	60–69	0.1%	14.8%	65.5%	19.5%
	70 +	0.0%	0.0%	6.0%	94.0%
MATH	00–49	93.2%	7.7%	0.1%	0.0%
	50–59	16.0%	64.0%	19.7%	0.2%
	60–69	0.1%	14.9%	65.9%	19.1%
	70 +	0.0%	0.0%	7.7%	92.3%

The conditional standard errors near the decision ranges were between 4 and 5 points. Table 4.8 shows classification results for a decision point of 50 correct for students whose true scores are more than 5 points (about one standard error) below or above the decision point and for students within 5 points of the decision point. For students outside this standard-error range, 98 to 99% were correctly classified. This seems quite reasonable and is about as far as we can go with statistical analyses. Beyond this point lie policy judgments about the adequacy and consequences of different rates of decision errors.

Table 4.8 Number of Students Scoring Below/Above 50 by True Score Level

Subject	True (Expected) Number Correct	Number of Students	Percent of These Students Who Would Actually Score:	
			< 50	50+
ELA	00.00–44.99	28.5	98.0%	2.0%
	45.00–54.99	15.5	45.9%	54.1%
	55.00+	56.0	0.7%	99.3%
MATH	00.00–44.99	47.0	98.5%	1.5%
	45.00–54.99	17.7	47.7%	52.3%
	55.00+	35.3	0.9%	99.1%

Potential Adverse Impact

It is not surprising that average item passing rates varied by demographic group since other indicators of student achievement, most notably recent STAR results, vary similarly by demographic group. We would like to know how passing rates for the test, as a whole, will vary for these different groups, but at this point, passing standards have not yet been set for the HSEE. Ultimately, judgments about how many of the standards must be mastered and to what degree need to be translated into a minimum performance level for a set of test items.

To provide a basis for interpreting performance difference on HSEE items across various groups, we examined results from the 1999 administration of STAR. In the principal surveys described in Chapter 5, nearly half of the principals estimated that fewer than half of last year's 10th graders would meet the HSEE requirements. The 1999 STAR results include comparisons to National norms. We looked at the percent of students in key demographic groups scoring above the 25th percentile for the nation as a whole (meaning that 75% of the nation's students would pass). For ELA, the passing rate for California would be about 50%. We also looked at what would happen if much higher standards, corresponding to the 50th percentile for the nation as a whole, were implemented. Table 4.9 shows the results.

Table 4.9 Percent of Students Above National Norm Values by Demographic Group

	Reading		Mathematics	
Demographic Group	National 25th Percentile	National 50th Percentile	National 25th Percentile	National 50th Percentile
All California Students	55	33	70	44
ELL Students	13	3	50	20
African American	40	17	52	22
Hispanic	36	15	57	25
Low SES (Parents Ed < HS)	28	10	54	23

As suggested both by the results in Table 4.9 and the field test score statistics in Tables 4.2 and 4.3, there will almost certainly be lower passing rates for Hispanic, African American, English-language learners (ELL), and low socioeconomic (SES) students than for students in general. The potentially great negative impact of denying more minority, ELL, and low SES students a diploma should be carefully considered. If the program works as intended, this negative consequence would be offset by programs that reduced the number of these students, with or without a high school diploma, who fail to develop language arts and mathematical skills that are critical to success beyond high school.

Summary

Overall, the results from the HSEE field test were quite positive. Notwithstanding the long test length, nearly all the students answered all of the items. The sample sizes, while less than hoped for, were adequate to provide stable estimates of both traditional and IRT item parameters. One limitation was the relatively modest number of African American, special needs, and ELL students who were tested, making it difficult to determine whether the items functioned differently for these groups.

Relatively few items had obvious statistical problems. This result confirmed results of direct observation that the item development and review process was thorough and effective.

Efforts to examine the potential accuracy of the HSEE scores, while very preliminary, were also reasonably positive. Even if the decision point were near the middle of the achievement distribution, 90% of the students taking an ELA form and 84% of the students taking a Math form would score more than one standard error above or below the decision point and these students would be classified (passed or failed) correctly at least 98% of the time.

One concern raised by the field test results was the relative difficulty of the items, particularly in mathematics. If these items reflect what we believe students need to know and be able to do, and several panels of reviewers believe that they do, then a significant number of 10th grade students are likely to fail this exam. Groups who traditionally score lower on assessments of student achievement will fail at higher rates. It will be important, therefore, to ensure that there are effective programs to help students at risk, both before and after their initial experience with this exam. It is possible that students will perform at higher levels during operational testing than they did on this field test where the results do not count. However, the very high completion rates suggest that nearly all students took the field test seriously.